文章编号: 1006-4710(2011)03-0368-05

基于离散小波变换和 RBF 神经网络的说话人识别

杨凯峰[1], 牟莉[2], 许亮[3]

(1. 西安理工大学 计算机科学与工程学院, 陕西 西安 710048; 2. 西安工程大学 计算机科学学院, 陕西 西安 710048;
3. 西安交通大学 电子与信息工程学院, 陕西 西安 710049)

摘要: 为提高说话人识别系统的性能,结合离散小波变换与 RBF 神经网络提出一种说话人识别新方法。把小波变换与美尔频率倒谱系数提取相结合,使用离散小波变换代替美尔频率倒谱系数中的离散余弦变换,提取变换谱振幅作为特征参数。使用逼近能力、分类能力和学习速度均更优的 RBF 神经网络取代常用的 BP 网络,采用与输入样本相关的方法优化 RBF 网络初始权值选取。不同语音长度和信噪比的实验表明,系统识别率和鲁棒性均得到了提高。

关键词: 说话人识别;MFWC;RBF 神经网络;初始权值

中图分类号: TN912.34 **文献标志码:** A

Speaker Recognition Based on Discrete Wavelet Transform and RBF Neural Networks

YANG Kaifeng[1], MOU Li[2], XU Liang[3]

(1. Faculty of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China;
2. School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China;
3. School of Electronic & Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: This paper presents a novel method of the speaker recognition in combining the discrete wavelet transform with RBF neural network so as to improve the speaker recognition system performances. The wavelet transform and Mel Frequency Cepstrum Coefficient extraction are combined. After displacing the discrete cosine transform with the wavelet transform, the amplitudes of transformed spectrum are extracted as the feature parameters. The BP networks are displaced by the RBF neural networks, with superior studying speed, approaching and characterizing ability. The initial weights choosing of the RBF networks are optimized by using an approach correlating with the input samples. Different speech length and SNR experiments show that the system recognition rate and robustness are all improved.

Key words: speaker recognition; MFWC; RBF NN; initial weight

在说话人识别研究中,特征参数以及识别方法的优劣直接影响说话人识别的准确率[1-5]。在以往的研究中,主要利用基频轮廓(Pitch Contour, PC)[1]、线性预测倒谱系数(Linear Prediction Cepstrum Coefficient, LPCC)[6-8]、美尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)[9-11]等特征参数来进行说话人识别。这些特征参数都是基于短时平稳的假设条件,但语音信号是一种典型非平稳信号,其频谱特性随时间而改变。短时分析不能随信号的变化动态调整时频分辨率,它仅对语音的静态特征进行描述,忽略了语音的动态特征。在识别方法上,常用方法有矢量量化(Vector Quantization, VQ)[12-13]、混合高斯模型(Gaussian Mixture Model, GMM)[14-15]和人工神经网络(Artificial Neutral Networks, ANN)[2,16-17]等。由于人工神经网络具有自适应、自组织和自学习等优点,近年来在说话人识别中得到了广泛应用。但网络模型一般采用 BP 网络,而 BP 网络在训练时,权值调节采用的是

收稿日期: 2011-05-24

基金项目: 陕西省教育厅产业化基金资助项目(05JC13)。

作者简介: 杨凯峰(1971-),男,陕西西安人,讲师,研究方向为信息处理。E-mail:kaifyang@gmail.com。

负梯度下降法，这种调节权值的方法存在收敛速度慢和局部极小等局限性。

本研究把小波变换与美尔频率倒谱系数的提取相结合，使用离散小波变换代替美尔频率倒谱系数中的离散余弦变换，提取变换谱的振幅作为说话人识别的特征参数。在识别方法上使用 RBF 神经网络取代常用的 BP 网络。RBF 神经网络在逼近能力、分类能力和学习速度等方面均优于 BP 网络。实验和分析表明，系统的识别率和鲁棒性都得到了提高。

1 美尔频率离散小波变换系数

与基于线性预测的倒谱系数相比，美尔频率倒谱系数的优点是不依赖全极点语音产生模型的假定。它的抗频谱失真和抗噪声能力较强，并且考虑了人耳的听觉感知特性，从而提高了系统识别率。在进行传统的美尔频率倒谱系数提取时，采用的是离散余弦变换，考虑到小波变换的时频局部变换特性以及多分辨率解析优势，本研究使用离散小波变换取代传统的离散余弦变换，提取美尔频率离散小波变换系数（Mel Frequency Wavelet Coefficient，MFWC）作为特征参数。

MFWC 特征参数提取流程图如图 1 所示。首先对预处理后的信号加汉明窗，然后对信号进行快速傅里叶变换求的频谱，再求出频谱平方即功率谱，用 N 个 mel 带通滤波器对功率谱进行滤波。由于不同频带在人耳中叠加，因此需将每个滤波器频带内信号进行叠加。对每个滤波器输出取对数，从而得到对应频带的对数功率谱，然后进行离散小波变换，得到美尔频率离散小波变换系数。

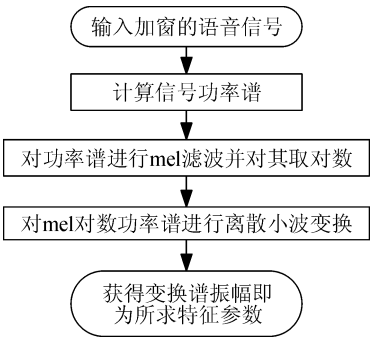

图 1 美尔频率离散小波变换系数提取流程图
Fig. 1 MFWC extraction flowchart

使用 Mallat 算法（金字塔算法）[18] 对待处理的语音信号进行离散小波变换。首先，离散平滑逼近递推由下式计算，为：

$$\boldsymbol{x}_k^{(j)} = \sum_n \boldsymbol{h}_0(n-2k) \times \boldsymbol{x}_k^{(j-1)} \tag{1}$$

离散细节信号递推表示为：

$$\boldsymbol{d}_k^{(j)} = \sum_n \boldsymbol{h}_1(n-2k) \times \boldsymbol{x}_k^{(j-1)} \tag{2}$$

其中，$\boldsymbol{h}_0(k)$ 和 $\boldsymbol{h}_1(k)$ 表示为：

$$\boldsymbol{h}_0(k) = \frac{1}{\sqrt{2}} \int \boldsymbol{\phi}\left(\frac{t}{2}\right) \boldsymbol{\phi}^*(t-k)\mathrm{d}t \tag{3}$$

$$\boldsymbol{h}_1(k) = \frac{1}{\sqrt{2}} \int \boldsymbol{\varphi}\left(\frac{t}{2}\right) \boldsymbol{\phi}^*(t-k)\mathrm{d}t \tag{4}$$

2 RBF 神经网络特征参数识别

人工神经网络对非线性映射具有出色的表达能力，1985 年，Powell 构造了多变量插值的 RBF 函数[19]。1988 年，Broomhead 和 Lowe 将插值计算演绎为神经计算，将 RBF 应用于人工神经网络设计，构造了 RBF 函数网络[20]。RBF 网络是一种三层前向神经网络，隐层激活函数为径向对称核函数。输入样本传播到隐单元空间时，这组核函数构成了输入样本的一组“基”。RBF 网络结构如图 2 所示。

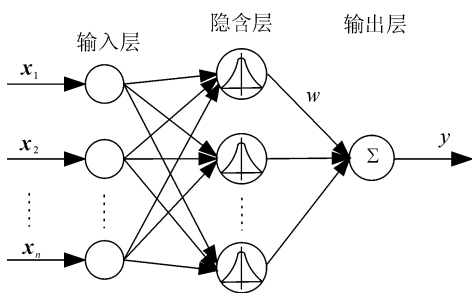

图 2 RBF 神经网络结构
Fig. 2 RBF neural networks structure

图 2 左端为输入层，完成将特征向量 $\boldsymbol{X}$ 引入网络。中间为隐含层，与输入层完全连接（权值为 1），其作用相当于对输入模式进行一次变换，将低维的模式输入数据变换到高维空间内，以利于输出层进行分类识别。隐层结点选取基函数作为转移函数，广泛使用的是高斯函数，即：

$$\boldsymbol{\Phi}_i(\boldsymbol{x}) = \exp[-\|\boldsymbol{x}-\boldsymbol{c}_i\|^2/(2\boldsymbol{\sigma}_i^2)]$$
$$i = 1,2,\cdots,p \tag{5}$$

式中，$\boldsymbol{x}$ 是 n 维输入向量；$\boldsymbol{c}_i$ 是第 i 个基函数的中心，与 $\boldsymbol{x}$ 具有相同维数的向量；$\boldsymbol{\sigma}_i$ 是第 i 个感知的变量，它决定了该基函数围绕中心点的宽度；p 是隐含层节点数。$\|\boldsymbol{x}-\boldsymbol{c}_i\|$ 是向量 $\boldsymbol{x}-\boldsymbol{c}_i$ 的范数，它表示 $\boldsymbol{x}$ 与 $\boldsymbol{c}_i$ 之间的距离；高斯函数在 $\boldsymbol{c}_i$ 处有一个唯一的最大值，随着 $\|\boldsymbol{x}-\boldsymbol{c}_i\|$ 的增大，函数逐渐衰减到零。应用 RBF 神经网络模型进行特征参数识别的流程图见图 3。

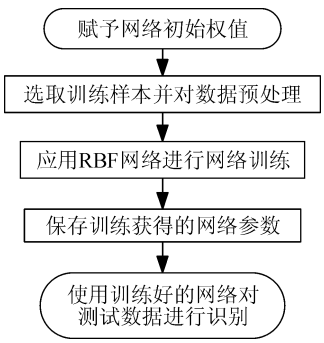

图3 RBF神经网络特征参数识别流程图
Fig. 3 RBF neural networks recognition flowchart

神经网络初始权值对的后续训练非常重要,好的初始权值能够加速网络训练的收敛速度,差的初始权值会导致学习次数增加,甚至不收敛。以往设置神经网络初始权值时,通常采用的方法是给其赋予一定范围内的随机数。虽然多数情况下这种方法能够取得收敛的训练结果,但缺陷在于初始权值与输入样本无相关性,对不同的样本难以获得最优的网络训练结果,因而为了获得好的结果往往要大量的重复试验。本研究采用与输入样本相关的权值选取方法来优化权值,这种初始权值选取方法即保证初始权值分布在样本的中心位置,又保证了其在样本中的离散性。具体步骤如下:

1) 对样本向量 $\boldsymbol{X}$ 进行归一化,即:

$$\hat{\boldsymbol{X}} = \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} = \left[\frac{\boldsymbol{x}_1}{\sqrt{\sum_{j=1}^{n}\boldsymbol{x}_j^2}},\cdots,\frac{\boldsymbol{x}_n}{\sqrt{\sum_{j=1}^{n}\boldsymbol{x}_j^2}}\right] \quad (6)$$

2) 通过样本分布计算权值分布,计算所有 $\hat{\boldsymbol{x}}_i$ 的平均值,即中心向量,然后计算每个向量与中心向量的欧氏距离,标记出最大距离 $d_{\max}$;

3) 假设样本的各属性之间独立,概率密度分布为 $f(\boldsymbol{x}) = \frac{1}{\boldsymbol{\sigma}\sqrt{2\pi}}\exp\left\{-\frac{\boldsymbol{x}-\boldsymbol{\mu}}{2\boldsymbol{\sigma}}\right\}$,$\boldsymbol{\mu}$ 为均值,$\boldsymbol{\sigma}$ 为标准差,以中心向量为中心点,构建多维正态分布,然后在 $d_{\max}$ 范围内选取随机数作为初始权值。

3 实验与分析

3.1 实验环境

实验采用 SoundMax 16 声卡采集语音数据,在安静环境下共录制了40人的语音,其中19名女性,21名男性。说话人对报刊、网络媒体上的文字随意朗读,对每个说话人分别录制10句作为其语音数据集。在每个说话人的语音数据集中,随机选取5句作为训练集,其余5句作为测试集。语音信号采样率为8 kHz,16 bits 量化,帧长为30 ms,帧移为10 ms,对信号进行预加重。

3.2 结果与分析

为了检验本研究方法的有效性,共进行了四组实验。首先在不同语音时长条件下,对纯净语音进行了不同特征参数与不同识别方法的两组实验;然后在相同语音时长条件下,对语音信号附加高斯白噪声,考察不同信噪比条件下系统识别率的变化。

实验一 不同语音长度下 MFWC 与 MFCC 特征参数识别性能比较。

分别采用 MFWC 和 MFCC 作为特征参数,识别方法均使用 RBF 神经网络,训练语音时长为 15 s,识别用语音时长分别为 1 s、3 s、6 s 和 9 s,识别率比较结果见图4。由图4可见,不同语音长度条件下,MFWC 比 MFCC 特征参数的系统识别率高。当语音长度增加时,说话人识别率提高,但变化幅度不大,这一结论复合物理规律。当语音长度从 9 s 下降到 1 s 时,MFWC 特征的识别率下降了 5.38%,而 MFCC 特征的识别率下降了 8.71%,说明 MFWC 特征比 MFCC 特征具有更强的鲁棒性。

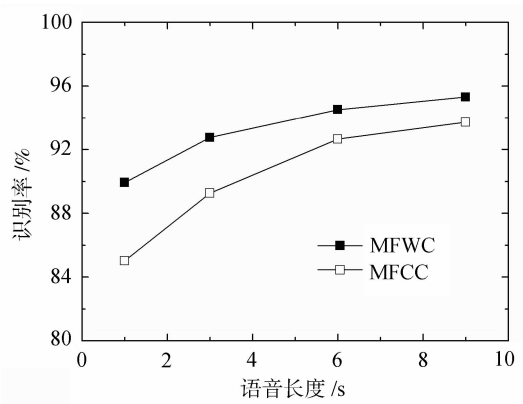

图4 不同语音长度 MFWC 与 MFCC 特征参数识别率比较
Fig. 4 Recognition rate comparison between MFWC and MFCC features at different speech lengths

实验二 不同语音长度下 RBF 神经网络与 BP 神经网络识别性能比较。

分别采用 RBF 神经网络与 BP 神经网络对说话人进行识别,在不改变实验一的实验条件下,特征参数均采用 MFWC,实验结果见图5。由图5可见,不同语音长度条件下,RBF 比 BP 神经网络的识别率高。当语音长度从 9 s 下降到 1 s 时,RBF 神经网络的识别率下降了 5.38%,而 BP 神经网络的识别率下降了 5.45%,这说明 RBF 神经网络识别的鲁棒性略高于 BP 神经网络识别。

实验三 不同信噪比条件下 MFWC 与 MFCC

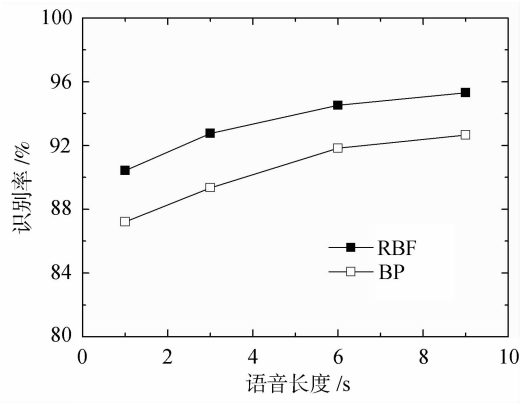

图 5 不同语音长度 RBF 与 BP 神经网络识别率比较
Fig. 5 Recognition rate comparison between RBF and BP neural networks at different speech lengths

特征参数识别率比较。

特征参数分别采用 MFWC 与 MFCC，识别方法采用 RBF 神经网络，训练语音时长为 15 s，识别语音时长为 3 s。分别加入信噪比为 5 dB、10 dB、15 dB和 20 dB 的白噪声，得到系统识别率比较结果见图 6。

由图 6 可见，在不同信噪比条件下，采用 MFWC 比采用 MFCC 特征参数具有更高的识别率，这一结论与实验一在不同语音长度条件下得到的结论相同。随着信噪比的增大，识别率也随之增加。当信噪比从 20 dB 下降到 5 dB 时，MFWC 特征参数的识别率下降了 14.49%，而 MFCC 特征参数的识别率下降了 17.41%。

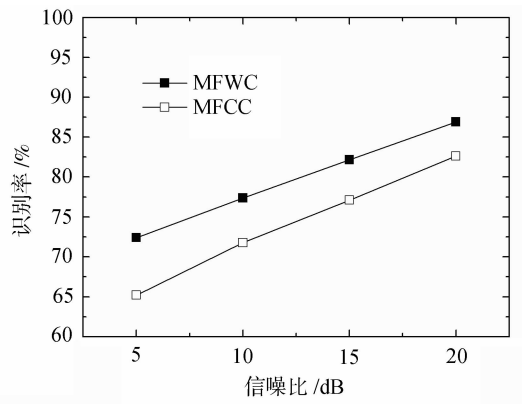

图 6 不同信噪比 MFWC 与 MFCC 特征参数识别率比较
Fig. 6 Recognition rate comparison between MFWC and MFCC features at different SNR

实验四 不同信噪比条件下 RBF 神经网络与 BP 神经网络识别性能比较。

识别方法分别采用 RBF 神经网络与 BP 神经网络，特征参数采用 MFWC，在不改变实验三的实验条件下，得到系统识别率比较结果，见图 7。

由图 7 可见，不同信噪比条件下，RBF 比采用 BP 神经网络具有更高的识别率，这一结论与实验二在不同语音长度条件下得到的结论相同。随着信噪比的增大，识别率增加。当信噪比从 20 dB 下降到 5 dB 时，RBF 神经网络的识别率下降了 14.49%，而 BP 神经网络的识别率下降了 15.26%。

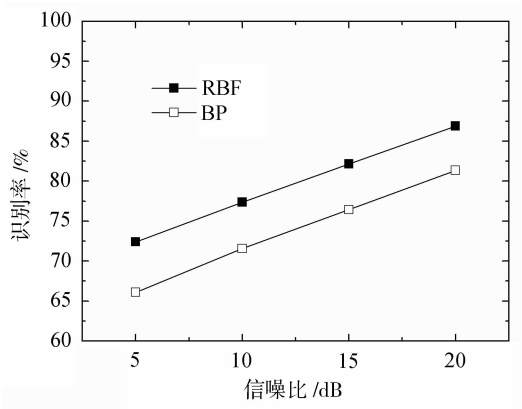

图 7 不同信噪比 RBF 与 BP 神经网络识别率比较
Fig. 7 Recognition rate comparison between RBF and BP neural networks at different SNR

4 结 论

本研究针对说话人识别系统中传统的 MFCC 特征参数提取和 BP 神经网络识别方法，使用分辨率可变且无平稳性要求的离散小波变换取代 MFCC 特征参数提取中的离散余弦变换，以及逼近能力、分类能力和学习速度均较优的 RBF 神经网络进行特征参数识别，使用与输入样本相关的方法来优化 RBF 网络初始权值选取。不同语音长度和信噪比的对比实验表明，采用离散小波变换的 MFWC 特征参数比 MFCC 特征参数具有更好的识别率和鲁棒性，RBF 神经网络识别比 BP 神经网络具有更高的识别率和鲁棒性。

参考文献：

[1] Bimbot F, Bonastre J F, Fredouille C, et al. A tutorial on text-independent speaker verification[J]. EURASIP Journal on Applied Signal Processing, 2004,(4): 430-451.

[2] Furui S. Digital Speech Processing, Synthesis, and Recognition[M]. New York: Marcel Dekker, 2000.

[3] Campbell W M, Campbell J P, Reynolds D A, et al. Support vector machines for speaker and language recognition [J]. Computer Speech and Language, 2006, 20(2): 210-229.

[4] Matsui T, Kanno T, Furui S. Speaker recognition using HMM composition in noisy environments[J]. Computer Speech and Language, 1996, 10(2): 107-116.

[5] Rabiner L R. A tutorial on hidden markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[6] Furui S. Fifty Years of Progress in Speech and Speaker Recognition: Proceedings of the 148th ASA Meeting[C]. San Diego: USA, 2004.

[7] Furui S. Cepstral analysis technique for automatic speaker verification[J]. IEEE Transactions on Acoustics, Speech, Signal , 1981, 29(2): 254-272.

[8] Furui S. Speaker independent isolated word recognition using dynamic features of speech spectrum[J]. IEEE Transactions on Acoustics, Speech, Signal Processing, 1986, 34(1): 52-59.

[9] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted gaussian mixture models[J]. Digital Signal , 2000, 10(1): 19-41.

[10] Reynolds D A. A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification [D]. Atlant: Georgia Institute of Technology, 1992.

[11] Reynolds D A, Rose R C. Robust text-independent speaker identification using gaussian mixture speaker models [J]. IEEE Transactions on Speech Audio , 1995, 3(1): 72-83.

[12] Rosenberg E, Soong F K. Evaluation of a vector quantization talker recognition system in text-independent and text-dependent models[J]. Computer Speech and Language, 1987, 2(3): 143-157.

[13] Soong F K, Rosenberg A. Rabiner L, et al. A vector quantization approach to speaker recognition[J] . AT&T Technical Journal, 1987, 66(1): 14-26.

[14] Rose R, Reynolds R A. Text Independent Speaker Identification Using Automatic Acoustic Segmentation: Proceedings of ICASSP[C]. Albuquerque: USA, 1990.

[15] Reynolds D A. Speaker identification and verification using gaussian mixture speaker models[J]. Speech Communication, 1995, 17(2): 91-108.

[16] Reynolds D A. An Overview of Automatic Speaker Recognition Technology: Proceedings of ICASSP[C]. Orlando: USA, 2002, 4072-4075.

[17] Katagiri S. Speech Pattern Recognition Using Neural Networks: Pattern Recognition in Speech and Language Processing[C]. Boca Raton: USA, 2003.

[18] 彭玉华,小波变换与工程应用[M]. 北京:科学出版社,2002.

[19] Power M J D. Radial Basis Function for Multivariable Interpolation: Proceedings of IMA Conference on Algorithms for the Approximation of Functions and Data [C]. Shrivenham: UK, 1985.

[20] Broomhead D S, Lowe D. Multivariable function interpolation and adaptive networks[J]. Complex System, 1988, 2: 321-355.

(责任编辑 李虹燕)